Near-Optimal Closeness Testing of Discrete Histogram Distributions

Near-Optimal Closeness Testing of Discrete Histogram Distributions*†

Ilias Diakonikolas[1], Daniel M. Kane[2], and Vladimir Nikishkin[3]

1 **University of Southern California, Los Angeles, CA, USA**
diakonik@usc.edu
2 **University of California, San Diego, CA, USA**
dakane@cs.ucsd.edu
2 **University of Edinburgh, Edinburgh, UK**
v.nikishkin@sms.ed.ac.uk

Abstract

We investigate the problem of testing the equivalence between two discrete histograms. A k-*histogram* over $[n]$ is a probability distribution that is piecewise constant over some set of k intervals over $[n]$. Histograms have been extensively studied in computer science and statistics. Given a set of samples from two k-histogram distributions p, q over $[n]$, we want to distinguish (with high probability) between the cases that $p = q$ and $\|p - q\|_1 \geq \epsilon$. The main contribution of this paper is a new algorithm for this testing problem and a nearly matching information-theoretic lower bound. Specifically, the sample complexity of our algorithm matches our lower bound up to a logarithmic factor, improving on previous work by polynomial factors in the relevant parameters. Our algorithmic approach applies in a more general setting and yields improved sample upper bounds for testing closeness of other structured distributions as well.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, G.3 Probability and Statistics

Keywords and phrases distribution testing, histograms, closeness testing

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.8

1 Introduction

In this work, we study the problem of testing equivalence (closeness) between two discrete *structured* distributions. Let $\mathcal{D}$ be a family of univariate distributions over $[n]$ (or $\mathbb{Z}$). The problem of *closeness testing for* $\mathcal{D}$ is the following: Given sample access to two unknown distribution $p, q \in \mathcal{D}$, we want to distinguish between the case that $p = q$ versus $\|p - q\|_1 \geq \epsilon$. (Here, $\|p-q\|_1$ denotes the ℓ_1-distance between the distributions p, q.) The sample complexity of this problem depends on the underlying family $\mathcal{D}$.

For example, if $\mathcal{D}$ is the class of *all* distributions over $[n]$, then it is known [13] that the optimal sample complexity is $\Theta(\max\{n^{2/3}/\epsilon^{4/3}, n^{1/2}/\epsilon^2\})$. This sample bound is best possible only if the family $\mathcal{D}$ includes all possible distributions over $[n]$, and we may be able to obtain significantly better upper bounds for most natural settings. For example, if both p, q are promised to be (approximately) log-concave over $[n]$, there is an algorithm to test

* A full version of this paper is available at https://arxiv.org/abs/1703.01913.
† I. D. was supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. D. K. was supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship. V. N. was supported by a University of Edinburgh PCD Scholarship.

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl;
Article No. 8; pp. 8:1–8:15


Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

equivalence between them using $O(1/\epsilon^{9/4})$ samples [25]. This sample bound is independent of the support size n, and is dramatically better than the worst-case tight bound [13] when n is large.

More generally, [25] described a framework to obtain sample-efficient equivalence testers for various families of structured distributions over both continuous and discrete domains. While the results of [25] are sample-optimal for *some* families of distributions (in particular, over continuous domains), it was not known whether they can be improved for natural families of discrete distributions. In this paper, we work in the framework of [25] and obtain new nearly-matching algorithms and lower bounds.

Before we state our results in full generality, we describe in detail a concrete application of our techniques to the case of *histograms* – a well-studied family of structured discrete distributions with a plethora of applications.

Testing Closeness of Histograms. A *k-histogram* over $[n]$ is a probability distribution $p : [n] \to [0, 1]$ that is piecewise constant over some set of k intervals over $[n]$. The algorithmic difficulty in testing properties of such distributions lies in the fact that the location and "size" of these intervals is a priori unknown. Histograms have been extensively studied in statistics and computer science. In the database community, histograms [37, 14, 46, 33, 35, 36, 1] constitute the most common tool for the succinct approximation of data. In statistics, many methods have been proposed to estimate histogram distributions [44, 32, 45, 40, 21, 48, 38] in a variety of settings.

In recent years, histogram distributions have attracted renewed interested from the theoretical computer science community in the context of learning [18, 10, 11, 12, 23, 2, 3, 27] and testing [36, 17, 26, 8, 9]. Here we study the following testing problem: Given sample access to two distributions p, q over $[n]$ that are promised to be (approximately) k-histograms, distinguish between the cases that $p = q$ versus $\|p - q\|_1 \geq \epsilon$. As the main application of our techniques, we give a new testing algorithm and a nearly-matching information-theoretic lower bound for this problem.

We now provide a summary of previous work on this problem followed by a description of our new upper and lower bounds. We want to ϵ-test closeness in ℓ_1-distance between two k-histograms over $[n]$, where $k \leq n$. Our goal is to understand the optimal sample complexity of this problem as a function of $k, n, 1/\epsilon$. Previous work is summarized as follows:

- In [25], the authors gave a closeness tester with sample complexity $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$.
- The best known sample lower bound is $\Omega(\max\{k^{2/3}/\epsilon^{4/3}, k^{1/2}/\epsilon^2\})$. This straightforwardly follows from [13], since k-histograms can simulate *any* support k distribution.

Notably, none of the two bounds depends on the domain size n. Observe that the upper bound of $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ cannot be tight for the entire range of parameters. For example, for $n = O(k)$, the algorithm of [13] for testing closeness between arbitrary support n distributions has sample size $O(\max\{k^{2/3}/\epsilon^{4/3}, k^{1/2}/\epsilon^2\})$, matching the above sample complexity lower bound, up to a constant factor.

This simple example might suggest that the $\Omega(\max\{k^{2/3}/\epsilon^{4/3}, k^{1/2}/\epsilon^2\})$ lower bound is tight in general. We prove that this is not the case. The main conceptual message of our new upper bound and nearly-matching lower bound is the following:

> *The sample complexity of ϵ-testing closeness between two k-histograms over $[n]$ depends in a subtle way on the relation between the relevant parameters k, n and $1/\epsilon$.*

We find this fact rather surprising because such a phenomenon does *not* occur for the sample complexities of closely related problems. Specifically, testing the identity of a k-histogram

over $[n]$ to a *fixed* distribution has sample complexity $\Theta(k^{1/2}/\epsilon^2)$ [26]; and learning a k-histogram over $[n]$ has sample complexity $\Theta(k/\epsilon^2)$ [11]. Note that both these sample bounds are independent of n and are known to be tight for the entire range of parameters $k, n, 1/\epsilon$.

Our main positive result is a new closeness testing algorithm for k-histograms over $[n]$ with sample complexity $O\big(k^{2/3} \cdot \log^{4/3}(2 + n/k) \log(k)/\epsilon^{4/3}\big)$. Combined with the known upper bound of [25], we obtain the sample upper bound of

$$O\big(\max\big(\min(k^{4/5}/\epsilon^{6/5}, k^{2/3}\log^{4/3}(2+n/k)\log(k)/\epsilon^{4/3}), k^{1/2}\log^2(k)\log\log(k)/\epsilon^2\big)\big).$$

As our main negative result, we prove a lower bound of $\Omega(\min(k^{2/3}\log^{1/3}(2+n/k)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5}))$. The first term in this expression shows that the "$\log(2 + n/k)$" factor that appears in the sample complexity of our upper bound is in fact necessary, up to a constant power. In summary, these bounds provide a nearly-tight characterization of the sample complexity of our histogram testing problem for the entire range of parameters.

A few observations are in order to interpret the above bounds:

- When n goes to infinity, the $O(k^{4/5}/\epsilon^{6/5})$ upper bound of [25] is tight for k-histograms.
- When $n = \mathrm{poly}(k)$ and ϵ is not too small (so that the $k^{1/2}/\epsilon^2$ term does not kick in), then the right answer for the sample complexity of our problem is $(k^{2/3}/\epsilon^{4/3})\mathrm{polylog}(k)$.
- The terms "$k^{4/5}/\epsilon^{6/5}$" and "$k^{2/3}\log^{4/3}(2 + n/k)\log(k)/\epsilon^{4/3}$" appearing in the sample complexity become equal when n is exponential in k. Therefore, our new algorithm has better sample complexity than that of [25] for all $n \leq 2^{O(k)}$.

In the following subsection, we state our results in a general setting and explain how the aforementioned applications are obtained from them.

1.1 Our Results and Comparison to Prior Work

For a given family $\mathcal{D}$ of discrete distributions over $[n]$, we are interested in designing a closeness tester for distributions in $\mathcal{D}$. We work in the general framework introduced by [26, 25]. Instead of designing a different tester for any given family $\mathcal{D}$, the approach of [26, 25] proceeds by designing a generic equivalence tester under a *different metric* than the ℓ_1-distance. This metric, termed $\mathcal{A}_k$-distance [20], where $k \geq 2$ is a positive integer, interpolates between Kolmogorov distance (when $k = 2$) and the ℓ_1-distance (when $k = n$). It turns out that, for a range of structured distribution families $\mathcal{D}$, the $\mathcal{A}_k$-distance can be used as a proxy for the ℓ_1-distance for a value of $k \ll n$ [11]. For example, if $\mathcal{D}$ is the family of k-histograms over $[n]$, the $\mathcal{A}_{2k}$ distance between them is tantamount to their ℓ_1 distance. We can thus obtain an ℓ_1 closeness tester for $\mathcal{D}$ by plugging in the right value of k in a general $\mathcal{A}_k$ closeness tester.

To formally state our results, we will need some terminology.

Notation. We will use p, q to denote the probability mass functions of our distributions. If p is discrete over support $[n] := \{1, \ldots, n\}$, we denote by p_i the probability of element i in the distribution. For two discrete distributions p, q, their ℓ_1 and ℓ_2 distances are $\|p-q\|_1 = \sum_{i=1}^n |p_i - q_i|$ and $\|p-q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$. Fix a partition of the domain I into disjoint intervals $\mathcal{I} := (I_i)_{i=1}^{\ell}$. For such a partition $\mathcal{I}$, the *reduced distribution* $p_r^{\mathcal{I}}$ corresponding to p and $\mathcal{I}$ is the discrete distribution over $[\ell]$ that assigns the i-th "point" the mass that p assigns to the interval I_i; i.e., for $i \in [\ell]$, $p_r^{\mathcal{I}}(i) = p(I_i)$. Let $\mathfrak{J}_k$ be the collection of all partitions of the domain I into k intervals. For $p, q : I \to \mathbb{R}_+$ and $k \in \mathbb{Z}_+$, we define the $\mathcal{A}_k$-distance between p and q by $\|p-q\|_{\mathcal{A}_k} \stackrel{\mathrm{def}}{=} \max_{\mathcal{I}=(I_i)_{i=1}^k \in \mathfrak{J}_k} \sum_{i=1}^k |p(I_i) - q(I_i)| = \max_{\mathcal{I}\in\mathfrak{J}_k} \|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_1$.

In this context, [25] gave a closeness testing algorithm under the $\mathcal{A}_k$-distance using $O(\max\{k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2\})$ samples. It was also shown that this sample bound is information–theoretically optimal (up to constant factors) for some adversarially constructed continuous distributions, or discrete distributions of support size n sufficiently large as a function of k. These results raised two natural questions: (1) What is the *optimal* sample complexity of the $\mathcal{A}_k$-closeness testing problem as a function of $n, k, 1/\epsilon$? (2) Can we obtain tight sample lower bounds for *natural* families of structured distributions?

We resolve both these open questions. Our main algorithmic result is the following:

▶ **Theorem 1.** *Given sample access to distributions p and q on $[n]$ and $\epsilon > 0$ there exists an algorithm that takes*

$$O\left(\max\left(\min\left(k^{4/5}/\epsilon^{6/5}, k^{2/3}\log^{4/3}(2+n/k)\log(2+k)/\epsilon^{4/3}\right), k^{1/2}\log^2(k)\log\log(k)/\epsilon^2\right)\right)$$

samples from each of p and q and distinguishes with $2/3$ probability between the cases that $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \epsilon$.

As explained in [26, 25], using Theorem 1 one can obtain testing algorithms for the ℓ_1 closeness testing of various distribution families $\mathcal{D}$, by using the $\mathcal{A}_k$ distance as a "proxy" for the ℓ_1 distance:

▶ **Fact 2.** *For a univariate distribution family $\mathcal{D}$ and $\epsilon > 0$, let $k = k(\mathcal{D}, \epsilon)$ be the smallest integer such that for any $f_1, f_2 \in \mathcal{D}$ it holds that $\|f_1 - f_2\|_1 \leq \|f_1 - f_2\|_{\mathcal{A}_k} + \epsilon/2$. Then there exists an ℓ_1 closeness testing algorithm for $\mathcal{D}$ with the sample complexity of Theorem 1.*

Applications

Our upper bound for ℓ_1-testing of k-histogram distributions follows from the above by noting that for any k-histograms p, q we have $\|p-q\|_1 = \|p-q\|_{\mathcal{A}_{2k}}$. Also note that our upper bound is *robust*: it applies even if p, q are $O(\epsilon)$-close in ℓ_1-norm to being k-histograms.

Finally, we remark that our general $\mathcal{A}_k$ closeness tester yields improved upper bounds for various other families of structured distributions. Consider for example the case that $\mathcal{D}$ consists of all k-mixtures of some simple family (e.g., discrete Gaussians or log-concave), where the parameter k is large. The algorithm of [25] leads to a tester whose sample complexity scales with $O(k^{4/5})$, while Theorem 1 implies a $\tilde{O}(k^{2/3})$ bound.

On the lower bound side, we show:

▶ **Theorem 3.** *Let p and q be distributions on $[n]$ and let $\epsilon > 0$ be less than a sufficiently small constant. Any tester that distinguishes between $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \epsilon$ for some $k \leq n$ must use $\Omega(m)$ samples for $m = \min(k^{2/3}\log^{4/3}(2+n/k)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$.*

Furthermore, for $m = \min(k^{2/3}\log^{1/3}(2+n/k)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$, any tester that distinguishes between $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \epsilon$ must use $\Omega(m)$ samples even if p and q are both guaranteed to be piecewise constant distributions on $O(k+m)$ pieces.

Note that a lower bound of $\Omega(\sqrt{k}/\epsilon^2)$ straightforwardly applies even for p and q being k-histograms. This dominates the above bounds for $\epsilon < k^{-3/8}$.

We also note that our general lower bound with respect to the $\mathcal{A}_k$ distance is somewhat stronger, matching the term "$\log^{4/3}(2+n/k)$" in our upper bound.

1.2 Related Work

During the past two decades, *distribution property testing* [5] – whose roots lie in statistical hypothesis testing [41, 39] – has received considerable attention by the computer science community, see [43, 7] for two recent surveys. The majority of the early work in this field has focused on characterizing the sample size needed to test properties of arbitrary distributions of a given support size. After two decades of study, this "worst-case" regime is well-understood: for many properties of interest there exist sample-optimal testers (matched by information-theoretic lower bounds) [42, 13, 47, 26, 24, 22].

In many settings of interest, we know a priori that the underlying distributions have some "nice structure" (exactly or approximately). The problem of *learning* a probability distribution under such structural assumptions is a classical topic in statistics, see [4] for a classical book, and [34] for a recent book on the topic, that has recently attracted the interest of computer scientists [18, 19, 10, 16, 11, 12, 1, 30, 31, 28, 15, 3, 27, 29].

On the other hand, the theory of *distribution testing* under structural assumptions is less fully developed. More than a decade ago, Batu, Kumar, and Rubinfeld [6] considered a specific instantiation of this question – testing the equivalence between two unknown discrete monotone distributions – and obtained a tester whose sample complexity is poly-logarithmic in the domain size. A recent sequence of works [17, 26, 25] developed a framework to leverage such structural assumptions and obtained more efficient testers for a number of natural settings. However, for several natural properties of interest there is still a substantial gap between known sample upper and lower bounds.

1.3 Overview of Techniques

To prove our upper bound, we use a technique of iteratively reducing the number of bins (domain elements). In particular, we show that if we merge bins together in consecutive pairs, this does not significantly affect the $\mathcal{A}_k$ distance between the distributions, unless a large fraction of the discrepancy between our distributions is supported on $O(k)$ bins near the boundaries in the optimal partition. In order to take advantage of this, we provide a novel identity tester that requires few samples to distinguish between the cases where $p = q$ and the case where p and q have a large ℓ_1 distance supported on only k of the bins. We are able to take advantage of the small support essentially because having a discrepancy supported on few bins implies that the ℓ_2 distance between the distributions must be reasonably large.

Our new lower bounds are somewhat more involved. We prove them by exhibiting explicit families of pairs of distributions, where in one case $p = q$ and in the other p and q have large $\mathcal{A}_k$ distance, but so that it is information-theoretically impossible to distinguish between these two families with a small number of samples. In both cases, p and q are explicit piecewise constant distributions with a small number of pieces. In both cases, our domain is partitioned into a small number of bins and the restrictions of the distributions to different bins are independent, making our analysis easier. In some bins we will have $p = q$ each with mass about $1/m$ (where m is the number of samples). These bins will serve the purpose of adding "noise" making harder to read the "signal" from the other bins. In the remaining bins, we will have either that $p = q$ being supported on some interval, or p and q will be supported on consecutive, non-overlapping intervals. If three samples are obtained from any one of these intervals, the order of the samples and the distributions that they come from will provide us with information about which family we came from. Unfortunately, since triple collisions are relatively uncommon, this will not be useful unless $m \gg \max(k^{4/5}/\epsilon^{6/5}, k^{1/2}/\epsilon^2)$. Bins from which we have one or zero samples will tell us nothing, but bins from which we have exactly two samples may provide information.

For these bins, it can be seen that we learn nothing from the ordering of the samples, but we may learn something from their spacing. In particular, in the case where p and q are supported on disjoint intervals, we would suspect that two samples very close to each other are far more likely to be taken from the same distribution rather than from opposite distributions. On the other hand, in order to properly interpret this information, we will need to know something about the scale of the distributions involved in order to know when two points should be considered to be "close". To overcome this difficulty, we will stretch each of our distributions by a random exponential amount. This will effectively conceal any information about the scales involved so long as the total support size of our distributions is exponentially large.

2 A Near-Optimal Closeness Tester over Discrete Domains

2.1 Warmup: A Simpler Algorithm

We start by giving a simpler algorithm establishing a basic version of Theorem 1 with slightly worse parameters:

▶ **Proposition 4.** *Given sample access to distributions p and q on $[n]$ and $\epsilon > 0$ there exists an algorithm that takes*

$$O\left(k^{2/3}\log^{4/3}(3+n/k)\log\log(3+n/k)/\epsilon^{4/3} + \sqrt{k}\log^2(3+n/k)\log\log(3+n/k)/\epsilon^2\right)$$

samples from each of p and q and distinguishes with $2/3$ probability between the cases that $p = q$ and $\|p-q\|_{\mathcal{A}_k} \geq \epsilon$.

The basic idea of our algorithm is the following: From the distributions p and q construct new distributions p' and q' by merging pairs of consecutive buckets. Note that p' and q' each have much smaller domains (of size about $n/2$). Furthermore, note that the $\mathcal{A}_k$ distance between p and q is $\sum_{I\in\mathcal{I}}|p(I)-q(I)|$ for some partition $\mathcal{I}$ into k intervals. By using essentially the same partition, we can show that $\|p'-q'\|_{\mathcal{A}_k}$ should be almost as large as $\|p-q\|_{\mathcal{A}_k}$. This will in fact hold unless much of the error between p and q is supported at points near the endpoints of intervals in $\mathcal{I}$. If this is the case, it turns out there is an easy algorithm to detect this discrepancy. We require the following definitions:

▶ **Definition 5.** For a discrete distribution p on $[n]$, the merged distribution obtained from p is the distribution p' on $\lceil n/2\rceil$, so that $p'(i) \stackrel{\text{def}}{=} p(2i)+p(2i+1)$. For a partition $\mathcal{I}$ of $[n]$, define the *divided partition* $\mathcal{I}'$ of domain $\lceil n/2\rceil$, so that $I'_i \in \mathcal{I}'$ has the points obtained by point-wise gluing together odd points and even points.

Note that one can simulate a sample from p' given a sample from p by letting $p' = \lceil p/2\rceil$.

▶ **Definition 6.** Let p and q be distributions on $[n]$. For integers $k \geq 1$, let $\|p-q\|_{1,k}$ be the sum of the largest k values of $|p(i)-q(i)|$ over $i \in [n]$.

We begin by showing that either $\|p'-q'\|_{\mathcal{A}_k}$ is close to $\|p-q\|_{\mathcal{A}_k}$ or $\|p-q\|_{1,k}$ is large.

▶ **Lemma 7.** *For any two distributions p and q on $[n]$, let p' and q' be the merged distributions. Then,*

$$\|p-q\|_{\mathcal{A}_k} \leq \|p'-q'\|_{\mathcal{A}_k} + 2\|p-q\|_{1,k}\ .$$

Proof. Let $\mathcal{I}$ be the partition of $[n]$ into k intervals so that $\|p-q\|_{\mathcal{A}_k} = \sum_{I\in\mathcal{I}} |p(I)-q(I)|$. Let $\mathcal{I}'$ be obtained from $\mathcal{I}$ by rounding each upper endpoint of each interval except for the last down to the nearest even integer, and rounding the lower endpoint of each interval up to the nearest odd integer. Note that

$$\sum_{I\in\mathcal{I}'} |p(I)-q(I)| = \sum_{I\in\mathcal{I}'} |p'(I/2)-q'(I/2)| \le \|p'-q'\|_{\mathcal{A}_k} .$$

The partition $\mathcal{I}'$ is obtained from $\mathcal{I}$ by taking at most k points and moving them from one interval to another. Therefore, the difference

$$\left| \sum_{I\in\mathcal{I}} |p(I)-q(I)| - \sum_{I\in\mathcal{I}'} |p(I)-q(I)| \right| ,$$

is at most twice the sum of $|p(i)-q(i)|$ over these k points, and therefore at most $2\|p-q\|_{1,k}$. Combing this with the above gives our result. ◀

Next, we need to show that if two distributions have $\|p-q\|_{1,k}$ large that this can be detected easily.

▶ **Lemma 8.** *Let p and q be distributions on $[n]$. Let $k>0$ be a positive integer, and $\epsilon>0$. There exists an algorithm which takes $O(k^{2/3}/\epsilon^{4/3}+\sqrt{k}/\epsilon^2)$ samples from each of p and q and, with probability at least $2/3$, distinguishes between the cases that $p=q$ and $\|p-q\|_{1,k}>\epsilon$.*

Note that if we needed to distinguish between $p=q$ and $\|p-q\|_1>\epsilon$, this would require $\Omega(n^{2/3}/\epsilon^{4/3}+\sqrt{n}/\epsilon^2)$ samples. However, the optimal testers for this problem are morally ℓ_2-testers. That is, roughly, they actually distinguish between $p=q$ and $\|p-q\|_2>\epsilon/\sqrt{n}$. From this viewpoint, it is clear why it would be easier to test for discrepancies in $\|-\|_{1,k}$-distance, since if $\|p-q\|_{1,k}>\epsilon$, then $\|p-q\|_2>\epsilon/\sqrt{k}$, making it easier for our ℓ_2-type tester to detect the difference.

Our general approach will be by way of the techniques developed in [24]. We begin by giving the definition of a split distribution coming from that paper:

▶ **Definition 9.** Given a distribution p on $[n]$ and a multiset S of elements of $[n]$, define the *split distribution* p_S on $[n+|S|]$ as follows: For $1\le i\le n$, let a_i denote 1 plus the number of elements of S that are equal to i. Thus, $\sum_{i=1}^n a_i = n+|S|$. We can therefore associate the elements of $[n+|S|]$ to elements of the set $B=\{(i,j): i\in[n], 1\le j\le a_i\}$. We now define a distribution p_S with support B, by letting a random sample from p_S be given by (i,j), where i is drawn randomly from p and j is drawn randomly from $[a_i]$.

We now recall two basic facts about split distributions:

▶ **Fact 10** ([24])**.** *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then:*

(i) *We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively.*

(ii) *It holds $\|p_S-q_S\|_1 = \|p-q\|_1$.*

▶ **Lemma 11** ([24])**.** *Let p be a distribution on $[n]$. Then:*

(i) *For any multisets $S\subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \le \|p_S\|_2$, and*

(ii) *If S is obtained by taking m samples from p, then $\mathbb{E}[\|p_S\|_2^2] \le 1/m$.*

We also recall an optimal ℓ_2 closeness tester under the promise that one of the distributions has smal ℓ_2 norm:

▶ **Lemma 12** ([13]). *Let p and q be two unknown distributions on $[n]$. There exists an algorithm that on input n, $b \geq \min\{\|p\|_2, \|q\|_2\}$ and $0 < \epsilon < \sqrt{2}b$, draws $O(b/\epsilon^2)$ samples from each of p and q and, with probability at least $2/3$, distinguishes between the cases that $p = q$ and $\|p - q\|_2 > \epsilon$.*

Proof of Lemma 8: We begin by presenting the algorithm:

Algorithm `Small-Support-Discrepancy-Tester`
Input: sample access to pdf's $p, q : [n] \to [0,1]$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{1,k} \geq \epsilon$.

1. Let $m = \min(k^{2/3}/\epsilon^{4/3}, k)$.
2. Let S be the multiset obtained by taking m independent samples from p.
3. Use the ℓ_2 tester of Lemma 12 to distinguish between the cases that $p_S = q_S$ and $\|p_S - q_S\|_2^2 \geq k^{-1}\epsilon^2/2$ and return the result.

The analysis is simple. By Lemma 11, with 90% probability $\|p_S\|_2 = O(m^{-1/2})$, and therefore the number of samples needed (using the ℓ_2 tester from Lemma 12) is $O(m + km^{-1/2}/\epsilon^2) = O(k^{2/3}/\epsilon^{4/3} + \sqrt{k}/\epsilon^2)$. If $p = q$, then $p_S = q_S$ and the algorithm will return "YES" with appropriate probability. If $\|q - p\|_{1,k} \geq \epsilon$, then $\|p_S - q_S\|_{1,k+m} \geq \epsilon$. Since $k + m$ elements contribute to total ℓ_1 error at least ϵ, by Cauchy-Schwarz, we have that $\|p_S - q_S\|_2^2 \geq \epsilon^2/(k+m) \geq k^{-1}\epsilon^2/2$. Therefore, in this case, the algorithm returns "NO" with appropriate probability. ◀

Proof of Proposition 4: The basic idea of our algorithm is the following: By Lemma 8, if $\|p - q\|_{\mathcal{A}_k}$ is large, then so is either $\|p - q\|_{1,k}$ or $\|p' - q'\|_{\mathcal{A}_k}$. Our algorithm then tests whether $\|p - q\|_{1,k}$ is large, and recursively tests whether $\|p' - q'\|_{\mathcal{A}_k}$ is large. Since p', q' have half the support size, we will only need to do this for $\log(n/k)$ rounds, losing only a poly-logarithmic factor in the sample complexity. We present the algorithm here:

Algorithm `Small-Domain-`$\mathcal{A}_k$`-tester`
Input: sample access to pdf's $p, q : [n] \to [0,1]$, $k \in \mathbb{Z}_+$, and $\epsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \epsilon$.

1. For $i := 0$ to $t \stackrel{\text{def}}{=} \lceil \log_2(n/k) \rceil$, let $p^{(i)}, q^{(i)}$ be distributions on $[\lceil 2^{-i} n \rceil]$ defined by $p^{(i)} = \lceil 2^{-i} p \rceil$ and $q^{(i)} = \lceil 2^{-i} q \rceil$.
2. Take $Ck^{2/3} \log^{4/3}(3 + n/k) \log\log(3 + n/k)/\epsilon^{4/3}$ samples, for C sufficiently large, and use these samples to distinguish between the cases $p^{(i)} = q^{(i)}$ and $\|p^{(i)} - q^{(i)}\|_{1,k} > \epsilon/(4\log_2(3+n/k))$ with probability of error at most $1/(10\log_2(3+n/k))$ for each i from 0 to t, using the same samples for each test.
3. If any test yields that $p^{(i)} \neq q^{(i)}$, return "NO". Otherwise, return "YES".

We now show correctness. In terms of sample complexity, we note that by taking a majority over $O(\log\log(3 + n/k))$ independent runs of the tester from Lemma 8 we can run this algorithm with the stated sample complexity. Taking a union bound, we can also assume that all tests performed in Step 2 returned the correct answer. If $p = q$ then $p^{(i)} = q^{(i)}$ for all i and thus, our algorithm returns "YES". Otherwise, we have that $\|p - q\|_{\mathcal{A}_k} \geq \epsilon$. By repeated application of Lemma 7, we have that

$$\|p - q\|_{\mathcal{A}_k} \leq \sum_{i=0}^{t-1} 2\|p^{(i)} - q^{(i)}\|_{1,k} + \|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} \leq 2\sum_{i=0}^{t} \|p^{(i)} - q^{(i)}\|_{1,k} \,,$$

where the last step was because $p^{(t)}$ and $q^{(t)}$ have a support of size at most k and so $\|p^{(t)} - q^{(t)}\|_{\mathcal{A}_k} = \|p^{(t)} - q^{(t)}\|_1 = \|p^{(t)} - q^{(t)}\|_{1,k}$. Therefore, if this is at least ϵ, it must be the case that $\|p^{(i)} - q^{(i)}\|_{1,k} > \epsilon/(4\log_2(3+n/k))$ for some $0 \leq i \leq t$, and thus our algorithm returns "NO". This completes our proof. ◀

2.2 Full Algorithm

The improvement to Proposition 4 is somewhat technical. The key idea involves looking into the analysis of Lemma 8. Generally speaking, choosing a larger value of m (up to the total sample complexity), will decrease the ℓ_2 norm of p, and thus the final complexity. Unfortunately, taking $m > k$ might lead to problems as it will subdivide the k original bins on which the error is supported into $\omega(k)$ bins. This in turn could worsen the lower bounds on $\|p - q\|_2$. However, this will only be the case if the total mass of these bins carrying the difference is large. Thus, we can obtain an improvement to Lemma 8 when the mass of bins on which the error is supported is small. The details are deferred to the full version.

3 Nearly Matching Information-Theoretic Lower Bound

We give a lower bound for k-histograms (k-flat distributions), postponing our slightly stronger construction to the full version. Before moving to the discrete setting, we first establish a lower bound for continuous histogram distributions. Our bound on discrete distributions will follow from taking the adversarial distribution from this example and rounding its values to the nearest integer. In order for this to work, we will need ensure to that our adversarial distribution does not have its $\mathcal{A}_k$-distance decrease by too much when we apply this operation. To satisfy this requirement, we will guarantee that our distributions will be piecewise constant with all the pieces of length at least 1.

▶ **Proposition 13.** *Let $k \in \mathbb{Z}_+$, $\epsilon > 0$ sufficiently small, and $W > 2$. Fix*

$$m = \min(k^{2/3}\log^{1/3}(W)/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5}) .$$

There exist distributions $\mathcal{D}, \mathcal{D}'$ over pairs of distributions p and q on $[0, 2(m+k)W]$, where p and q are $O(m+k)$-flat with pieces of length at least 1, so that: (a) when drawn from $\mathcal{D}$, we have $p = q$ deterministically, (b) when drawn from $\mathcal{D}'$, we have $\|p - q\|_{\mathcal{A}_k} > \epsilon$ with 90% probability, and so that $o(m)$ samples are insufficient to distinguish whether or not the pair is drawn from $\mathcal{D}$ or $\mathcal{D}'$ with better than 2/3 probability.

At a high-level, our lower bound construction proceeds as follows: We will divide our domain into $m+k$ bins so that no information about which distributions had samples drawn from a given bin or the ordering of these samples will help to distinguish between the cases of $p = q$ and otherwise, unless at least three samples are taken from the bin in question. Approximately k of these bins will each have mass ϵ/k and might convey this information if at least three samples are taken from the bin. However, the other m bins will each have mass approximately $1/m$ and will be used to add noise. In all, if we take s samples, we expect to see approximately $s^3\epsilon^3/k^2$ of the lighter bins with at least three samples. However, we will see approximately s^3/m^2 of our heavy bins with three samples. In order for the signal to overwhelm the noise, we will need to ensure that we have $(s^3\epsilon^3/k^2)^2 > s^3/m^2$.

The above intuitive sketch assumes that we cannot obtain information from the bins in which only two samples are drawn. This naively should not be the case. If $p = q$, the distance between two samples drawn from that bin will be independent of whether or not they are

drawn from the same distribution. However, if p and q are supported on disjoint intervals, one would expect that points that are close to each other should be far more likely to be drawn from the same distribution than from different distributions. In order to disguise this, we will scale the length of the intervals by a random, exponential amount, essentially making it impossible to determine what is meant by two points being close to each other. In effect, this will imply that two points drawn from the same bin will only reveal $O(1/\log(W))$ bits of information about whether $p = q$ or not. Thus, in order for this information to be sufficient, we will need that $(s^2\epsilon^2/k)^2/\log(W) > (s^2/m)$. We proceed with the formal proof below.

Proof of Proposition 13: We use ideas from [24] to obtain this lower bound using an information theoretic argument.

We may assume that $\epsilon > k^{1/2}$, because otherwise we may employ the standard lower bound that $\Omega(\sqrt{k}/\epsilon^2)$ samples are required to distinguish two distributions on a support of size k.

First, we note that it is sufficient to take $\mathcal{D}$ and $\mathcal{D}'$ be distributions over pairs of non-negative, piecewise constant distributions with total mass $\Theta(1)$ with 90% probability so that running a Poisson process with parameter $o(m)$ is insufficient to distinguish a pair from $\mathcal{D}$ from a pair from $\mathcal{D}'$ [24].

We construct these distributions as follows: We divide the domain into $m + k$ bins of length $2W$. For each bin i, we independently generate a random ℓ_i, so that $\log(\ell_i/2)$ is uniformly distributed over $[0, 2\log(W)/3]$. We then produce an interval I_i within bin i of total length ℓ_i and with random offset. In all cases, we will have p and q supported on the union of the I_i's.

For each i with probability $m/(m+k)$, we have the restrictions of p and q to I_i both uniform with $p(I_i) = q(I_i) = 1/m$. The other $k/(m + k)$ of the time we have $p(I_i) = q(I_i) = \epsilon/k$. In this latter case, if p and q are being drawn from $\mathcal{D}$, p and q are each constant on this interval. If they are being drawn from $\mathcal{D}'$, then $p + q$ will be constant on the interval, with all of that mass coming from p on a random half and coming from q on the other half.

Note that in all cases p and q are piecewise constant with $O(m + k)$ pieces of length at least 1. It is easy to show that with high probability the total mass of each of p and q is $\Theta(1)$, and that if drawn from $\mathcal{D}'$ that $\|p - q\|_{\mathcal{A}_k} \gg \epsilon$ with at least 90% probability.

We will now show that if one is given m samples from each of p and q, taken randomly from either $\mathcal{D}$ or $\mathcal{D}'$, that the shared information between the samples and the source family will be small. This implies that one is unable to consistently guess whether our pair was taken from $\mathcal{D}$ or $\mathcal{D}'$.

Let X be a random variable that is uniformly at random either 0 or 1. Let A be obtained by applying a Poisson process with parameter $s = o(m)$ on the pair of distributions p, q drawn from $\mathcal{D}$ if $X = 0$ or from $\mathcal{D}'$ if $X = 1$. We note that it suffices to show that the shared information $I(X : A) = o(1)$. In particular, by Fano's inequality, we have:

▶ **Lemma 14.** *If X is a uniform random bit and A is a correlated random variable, then if f is any function so that $f(A) = X$ with at least 51% probability, then $I(X : A) \geq 2 \cdot 10^{-4}$.*

Let A_i be the samples of A taken from the i^{th} bin. Note that the A_i are conditionally independent on X. Therefore, we have that $I(X : A) \leq \sum_i I(X : A_i) = (m + k)I(X : A_1)$. We will proceed to bound $I(X : A_1)$.

We note that $I(X : A_1)$ is at most the integral over pairs of multisets a (representing a set of samples from q and a set of samples from p), of

$$O\left(\frac{(\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1))^2}{\Pr(A_1 = a)}\right).$$

Thus,

$$I(X : A_1) = \sum_{h=0}^{\infty} \int_{|a|=h} O\left(\frac{(\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1))^2}{\Pr(A_1 = a)}\right).$$

We will split this sum up based on the value of h.

For $h = 0$, we note that the distributions for $p + q$ are the same for $X = 0$ and $X = 1$. Therefore, the probability of selecting no samples is the same. Therefore, this contributes 0 to the sum.

For $h = 1$, we note that the distributions for $p + q$ are the same in both cases, and conditioning on I_1 and $(p+q)(I_1)$ that $\mathbb{E}[p]$ and $\mathbb{E}[q]$ are the same in each of the cases $X = 0$ and $X = 1$. Therefore, again in this case, we have no contribution.

For $h \geq 3$, we note that $I(X : A_1) \leq I(X : A_1, I_1) \leq I(X : A_1|I_1)$, since I_1 is independent of X. We note that $\Pr(A_1 = a|X = 0, p(I_1) = 1/m) = \Pr(A_1 = a|X = 1, p(I_1) = 1/m)$. Therefore, we have that

$$\begin{aligned}\Pr(A_1 = a|X = 0) - \Pr(A_1 = a|X = 1)\\ = \Pr(A_1 = a|X = 0, p(I_1) = \epsilon/k) - \Pr(A_1 = a|X = 1, p(I_1) = \epsilon/k).\end{aligned}$$

If $p(I_1) = \epsilon/k$, the probability that exactly h elements are selected in this bin is at most $k/(m+k)(2s\epsilon/k)^h/h!$, and if they are selected, they are uniformly distributed in I_1 (although which of the sets p and q they are taken from is non-uniform). However, the probability that h elements are taken from I_1 is at least $\Omega(m/(m+k)(sm)^{-h}/h!)$ from the case where $p(I_1) = 1/m$, and in this case the elements are uniformly distributed in I_1 and uniformly from each of p and q. Therefore, we have that this contribution to our shared information is at most $k^2/(m(m+k))O(s\epsilon^2 m/k^2)^h/h!$. We note that $\epsilon^2 m/k^2 < 1$. Therefore, the sum of this over all $h \geq 3$ is $k^2/(m(m+k))O(s\epsilon^2 m/k^2)^3$. Summing over all $m + k$ bins, this is $k^{-4}\epsilon^6 s^3 m^2 = o(1)$.

It remains to analyze the case where $h = 2$. Once again, we have that ignoring which of p and q elements of A_1 came from, A_1 is identically distributed conditioned on $p(I_1) = 1/m$ and $|A_1| = 2$ as it is conditioned on $p(I_1) = \epsilon/k$ and $|A_1| = 2$. Since once again, the distributions $\mathcal{D}$ and $\mathcal{D}'$ are indistinguishable in the former case, we have that the contribution of the $h = 2$ terms to the shared information is at most

$$\begin{aligned}O\left(\frac{(k/(k+m)(\epsilon s/k)^2)^2}{m/(k+m)(s/m)^2}\right) d_{\mathrm{TV}}((A_1|X = 0, p(I_1)\epsilon/k, |A_1| = 2),\\ (A_1|X = 1, p(I_1) = \epsilon/k, |A_1| = 2))\end{aligned}$$

or

$$\begin{aligned}O\left(s^2 m k^{-2}\epsilon^4/(k+m)\right) d_{\mathrm{TV}}((A_1|X = 0, p(I_1) = \epsilon/k, |A_1| = 2),\\ (A_1|X = 1, p(I_1) = \epsilon/k, |A_1| = 2)) .\end{aligned}$$

It will suffice to show that conditioned upon $p(I_1) = \epsilon/k$ and $|A_1| = 2$ that

$$d_{\mathrm{TV}}((A_1|X = 0), (A_1|X = 1)) = O(1/\log(W)).$$

Let f be the order preserving linear function from $[0, 2]$ to I_1. Notice that conditional on $|A_1| = 2$ and $p(I_1) = \epsilon/k$ that we may sample from A_1 as follows:

- Pick two points $x > y$ uniformly at random from $[0, 2]$.
- Assign the points to p and q as follows:

- If $X = 0$ uniformly randomly assign these points to either distribution p or q.
- If $X = 1$ randomly do either:
  - Assign points in $[0, 1]$ to q and other points to p.
  - Assign points in $[0, 1]$ to p and other points to q.
- Randomly pick I_1 and apply f to x and y to get outputs $z = f(x), w = f(y)$.

Notice that the four cases:

(i) both points coming from p,
(ii) both points coming from q,
(iii) a point from p preceding a point from q,
(iv) a point from q preceding a point from p,

are all equally likely conditioned on either $X = 0$ or $X = 1$. However, we will note that this ordering is no longer independent of the choice of x and y.

Therefore, we can sample from A_1 subject to $X = 0$ and from A_1 subject to $X = 1$ in such a way that this ordering is the same deterministically. We consider running the above sampling algorithm to select (x, y) while sampling from $X = 0$ and (x', y') when sampling from $X = 1$ so that we are in the same one of the above four cases. We note that

$$d_{\mathrm{TV}}((A_1|X = 0), (A_1|X = 1)) \leq \mathbb{E}_{x,y,x',y'}[d_{\mathrm{TV}}((f(x), f(y)), (f(x'), f(y')))] \, ,$$

where the variation distance is over the random choices of f.

To show that this is small, we note that $|f(x) - f(y)|$ is distributed like $\ell_1(x - y)$. This means that $\log(|f(x) - f(y)|)$ is uniform over $[\log(f(x) - f(y)), \log(f(x) - f(y)) + 2\log(W)/3]$. Similarly, $\log(|f'(x') - f'(y')|)$ is uniform over $[\log(f(x') - f(y')), \log(f(x') - f(y')) + 2\log(W)/3]$. These differ in total variation distance by

$$O\left(\frac{|\log(f(x) - f(y))| + |\log(f(x') - f(y'))|}{\log(W)}\right) \, .$$

Taking the expectation over x, y, x', y' we get $O(1/\log(W))$. Therefore, we may further correlate the choices made in selecting our two samples, so that $z - w = z' - w'$ except with probability $O(1/\log(W))$. We note that after conditioning on this, z and z' are both uniformly distributed over subintervals of $[0, 2W]$ of length at least $2(W - W^{2/3})$. Therefore, the distributions on z and z' differ by at most $O(W^{-1/3})$. Hence, the total variation distance between A_1 conditioned on $|A_1| = 2, p(I_1) = \epsilon/k, X = 0$ and conditioned on $|A_1| = 2, p(I_1) = \epsilon/k, X = 1$ is at most $O(1/\log(W)) + O(W^{-1/3}) = O(1/\log(W))$. This completes our proof. ◀

We can now turn this into a lower bound for testing $\mathcal{A}_k$ distance on discrete domains.

Proof of second half of Theorem 3: Assume for sake of contradiction that this is not the case, and that there exists a tester taking $o(m)$ samples. We use this tester to come up with a continuous tester that violates Proposition 13.

We begin by proving a few technical bounds on the parameters involved. Firstly, note that we already have a lower bound of $\Omega(k^{1/2}/\epsilon^2)$, so we may assume that this is much less than m. We now claim that $m = O(\min(k^{2/3}\log^{1/3}(3 + n/(m + k))/\epsilon^{4/3}, k^{4/5}/\epsilon^{6/5})$. If $m \leq k$, there is nothing to prove. Otherwise,

$$k^{2/3}\log^{1/3}(3 + n/(m + k))/\epsilon^{4/3} \geq m(m/k)^{-1/3}\log(3 + n/(m + k))^{1/3}.$$

Thus, there is nothing more to prove unless $\log(3 + n/(m + k)) \gg m/k$. But, in this case, $\log(3 + n/(m + k)) \gg \log(m/k)$ and thus $\log(3 + n/(m + k)) = \Theta(\log(3 + n/k))$, and we are done.

We now let $W = n/(6(m+k))$, and let $\mathcal{D}$ and $\mathcal{D}'$ be as specified in Proposition 13. We claim that we have a tester to distinguish a p, q from $\mathcal{D}$ from ones taken from $\mathcal{D}'$ in $o(m)$ samples. We do this as follows: By rounding p and q down to the nearest third of an integer, we obtain p',q' supported on set of size n. Since p and q were piecewise constant on pieces of size at least 1, it is not hard to see that $\|p' - q'\|_{\mathcal{A}_k} \geq \|p - q\|_{\mathcal{A}_k}/3$. Therefore, a tester to distinguish $p' = q'$ from $\|p' - q'\|_{\mathcal{A}_k} \geq \epsilon$ can be used to distinguish $p = q$ from $\|p - q\|_{\mathcal{A}_k} \geq 3\epsilon$. This is a contradiction and proves our lower bound. ◀

References

1. J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. Fast and Near-Optimal Algorithms for Approximating Distributions by Histograms. In *34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2015*, pages 249–263, 2015.
2. J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Fast algorithms for segmented regression. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, pages 2878–2886, 2016.
3. J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1278–1289, 2017. Full version available at https://arxiv.org/abs/1506.00671.
4. R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.
5. T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000. URL: `citeseer.ist.psu.edu/batu00testing.html`.
6. T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.
7. C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.
8. C. L. Canonne. Are few bins enough: Testing histogram distributions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016*, pages 455–463, 2016.
9. C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016.
10. S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.
11. S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.
12. S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.
13. S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.
14. S. Chaudhuri, R. Motwani, and V. R. Narasayya. Random sampling for histogram construction: How much is enough? In *SIGMOD Conference*, pages 436–447, 1998.
15. C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC'16, New York, NY, USA, 2016. ACM.

16 C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

17 C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

18 C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

19 C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

20 L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

21 L. Devroye and G. Lugosi. Bin width selection in multivariate histograms by the combinatorial method. *Test*, 13(1):129–145, 2004.

22 I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016.

23 I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. In *NIPS*, pages 2566–2574, 2015.

24 I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

25 I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 1183–1202, 2015.

26 I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 1841–1854, 2015.

27 I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient robust proper learning of log-concave distributions. *CoRR*, abs/1606.03077, 2016.

28 I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of STOC'16*, 2016.

29 I. Diakonikolas, D. M. Kane, and A. Stewart. Learning multivariate log-concave distributions. *CoRR*, abs/1605.08188, 2016.

30 I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal Learning via the Fourier Transform for Sums of Independent Integer Random Variables. In *COLT*, volume 49, pages 831–849, 2016. Full version available at arXiv:1505.00662.

31 I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 850–878, 2016. Full version available at arXiv:1511.04066.

32 D. Freedman and P. Diaconis. On the histogram as a density estimator:l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981.

33 A. C. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *STOC*, pages 389–398, 2002.

34 P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

35 S. Guha, N. Koudas, and K. Shim. Approximation and streaming algorithms for histogram construction problems. *ACM Trans. Database Syst.*, 31(1):396–438, 2006.

36 P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *PODS*, pages 15–22, 2012.

37 H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. C. Sevcik, and T. Suel. Optimal histograms with quality guarantees. In *VLDB*, pages 275–286, 1998.

38 J. Klemela. Multivariate histograms with data-dependent partitions. *Statistica Sinica*, 19(1):159–176, 2009.

39 E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

40 G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*, 24(2):687–706, 04 1996.

41 J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi:10.1098/rsta.1933.0009.

42 L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

43 R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

44 D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

45 D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

46 N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. In *SIGMOD Conference*, pages 428–439, 2002.

47 G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

48 R. Willett and R. D. Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.